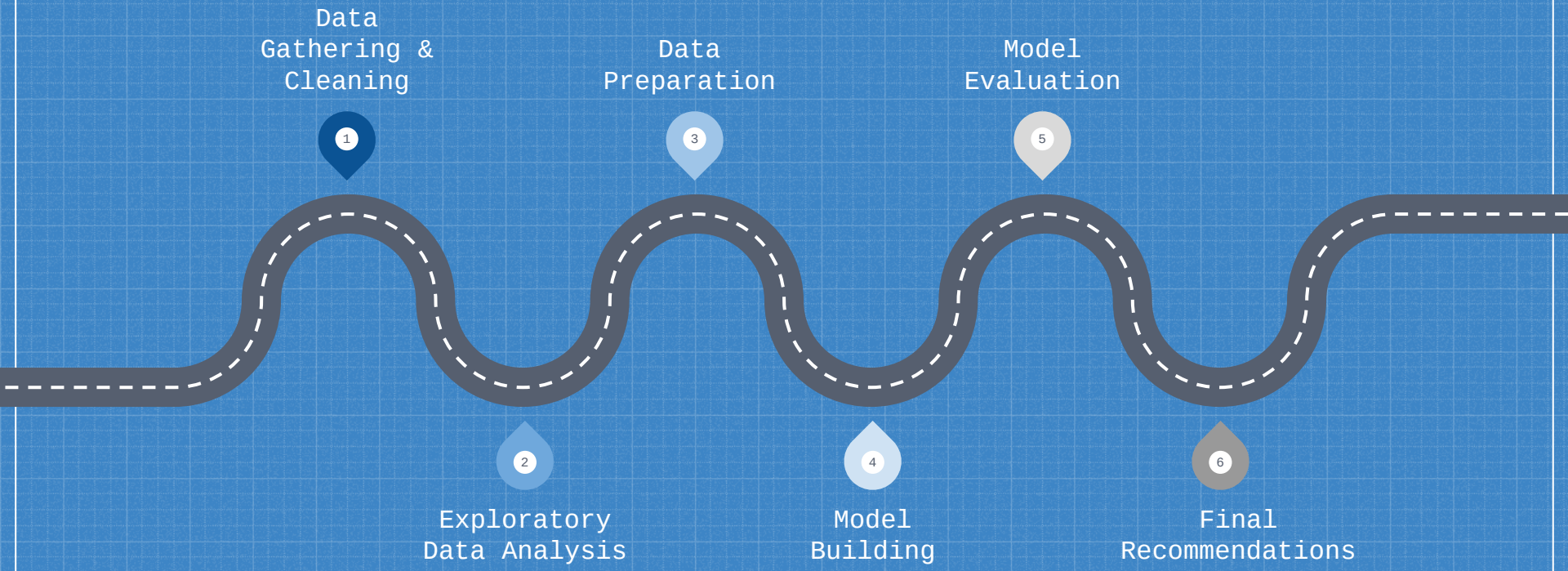
LEAD SCORING CASE STUDY

By
Gaurav Gilalkar
Vennela Surishetti

AGENDA

- Data Gathering and Cleaning
- Exploratory Data Analysis
- Data Preparation
- Model Building
- Model Evaluation
- Final recommendations

STEPS IN MODEL BUILDING



EXPLORATORY DATA ANALYSIS

DATA CLEANING

CHECKING FOR NULL VALUES

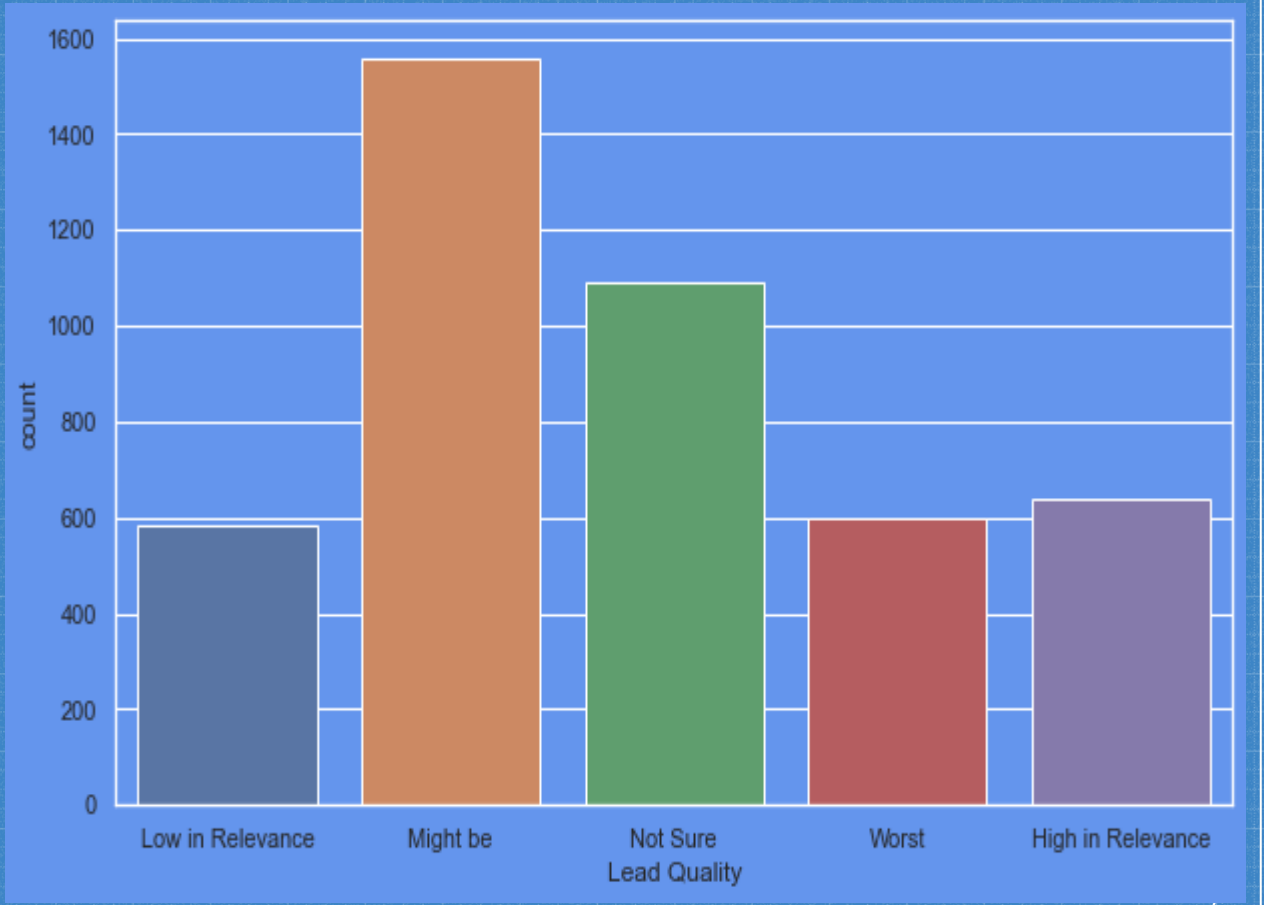
Columns having NULL values more than 70% will be dropped and the rest will be analyzed and imputed

```
# Checking the percentage of null values in all columns
round(100 * (leads.isnull().sum()/len(leads.index)), 2)
```

```
Prospect ID                                      0.00
Lead Number                                      0.00
Lead Origin                                      0.00
Lead Source                                      0.39
Do Not Email                                     0.00
Do Not Call                                      0.00
Converted                                        0.00
TotalVisits                                      1.48
Total Time Spent on Website                      0.00
Page Views Per Visit                             1.48
Last Activity                                    1.11
Country                                         26.63
Specialization                                  36.58
How did you hear about X Education              78.46
What is your current occupation                 29.11
What matters most to you in choosing a course   29.32
Search                                           0.00
Magazine                                         0.00
Newspaper Article                                0.00
X Education Forums                               0.00
Newspaper                                        0.00
Digital Advertisement                            0.00
Through Recommendations                          0.00
Receive More Updates About Our Courses           0.00
Tags                                            36.29
Lead Quality                                    51.59
Update me on Supply Chain Content                0.00
Get updates on DM Content                        0.00
Lead Profile                                    74.19
City                                            39.71
Asymmetrique Activity Index                     45.65
Asymmetrique Profile Index                      45.65
Asymmetrique Activity Score                     45.65
Asymmetrique Profile Score                      45.65
I agree to pay the amount through cheque         0.00
A free copy of Mastering The Interview           0.00
Last Notable Activity                            0.00
dtype: float64
```

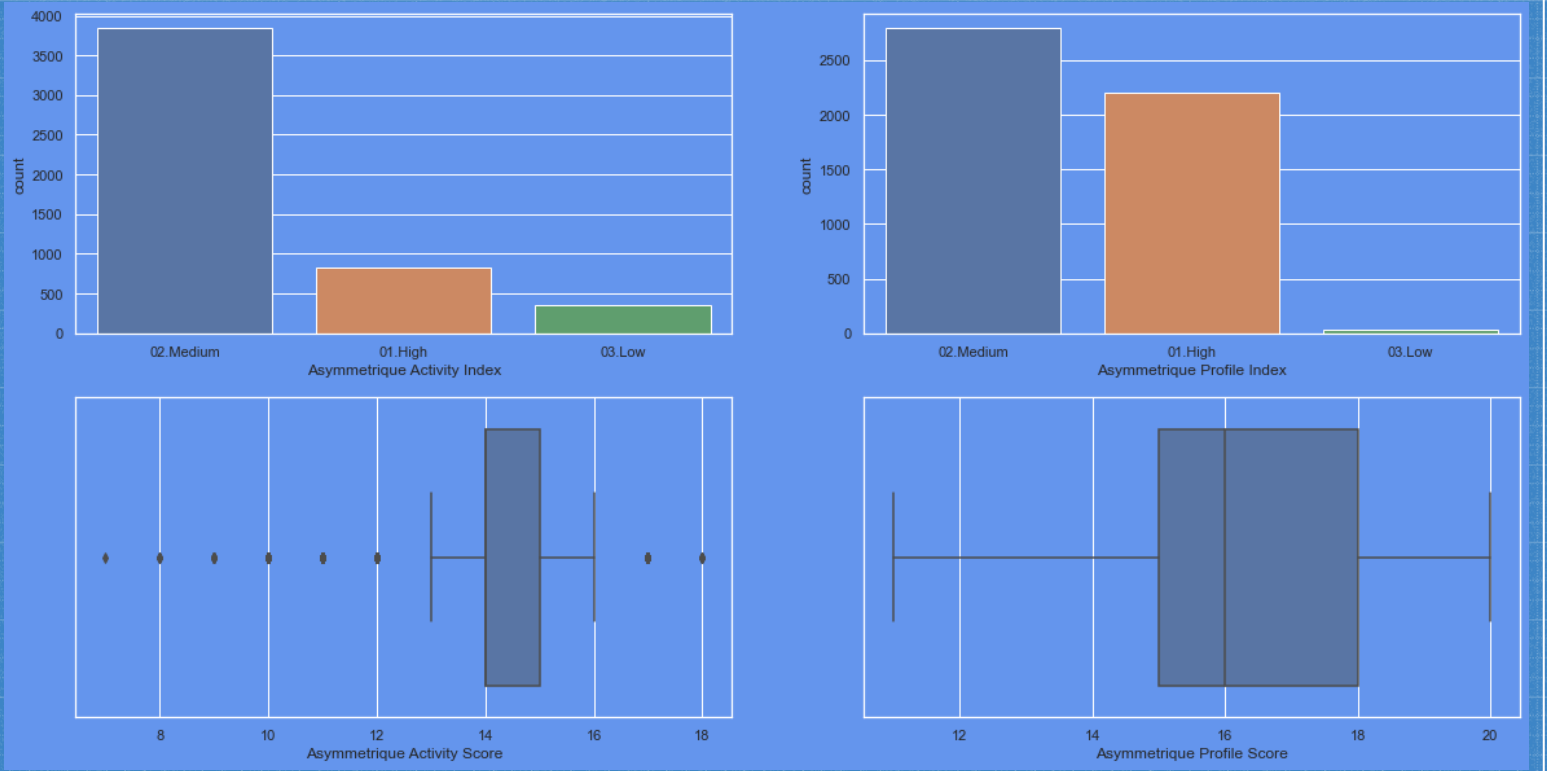
LEAD QUALITY

“Not Sure” seems to be the most neutral value so we replace the NULL values with “Not Sure”



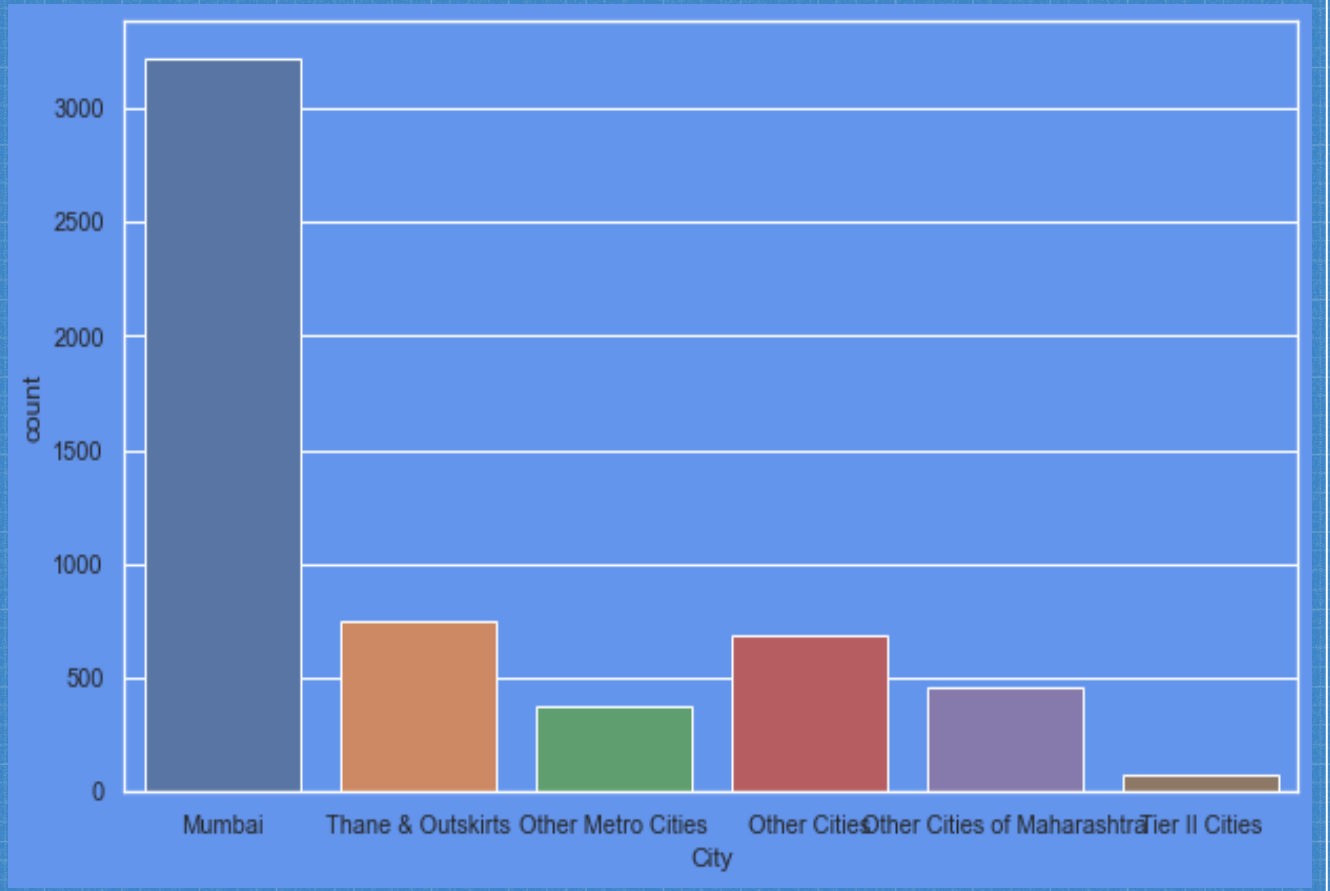
ASYMMETRIQUE ACTIVITY/ PROFILE SCORE/ INDEX

There is variation in data in these four columns and we were looking at the data in order to impute the NULL values (which are 45%) So, we can't make a conclusive decision on this, so we drop these columns



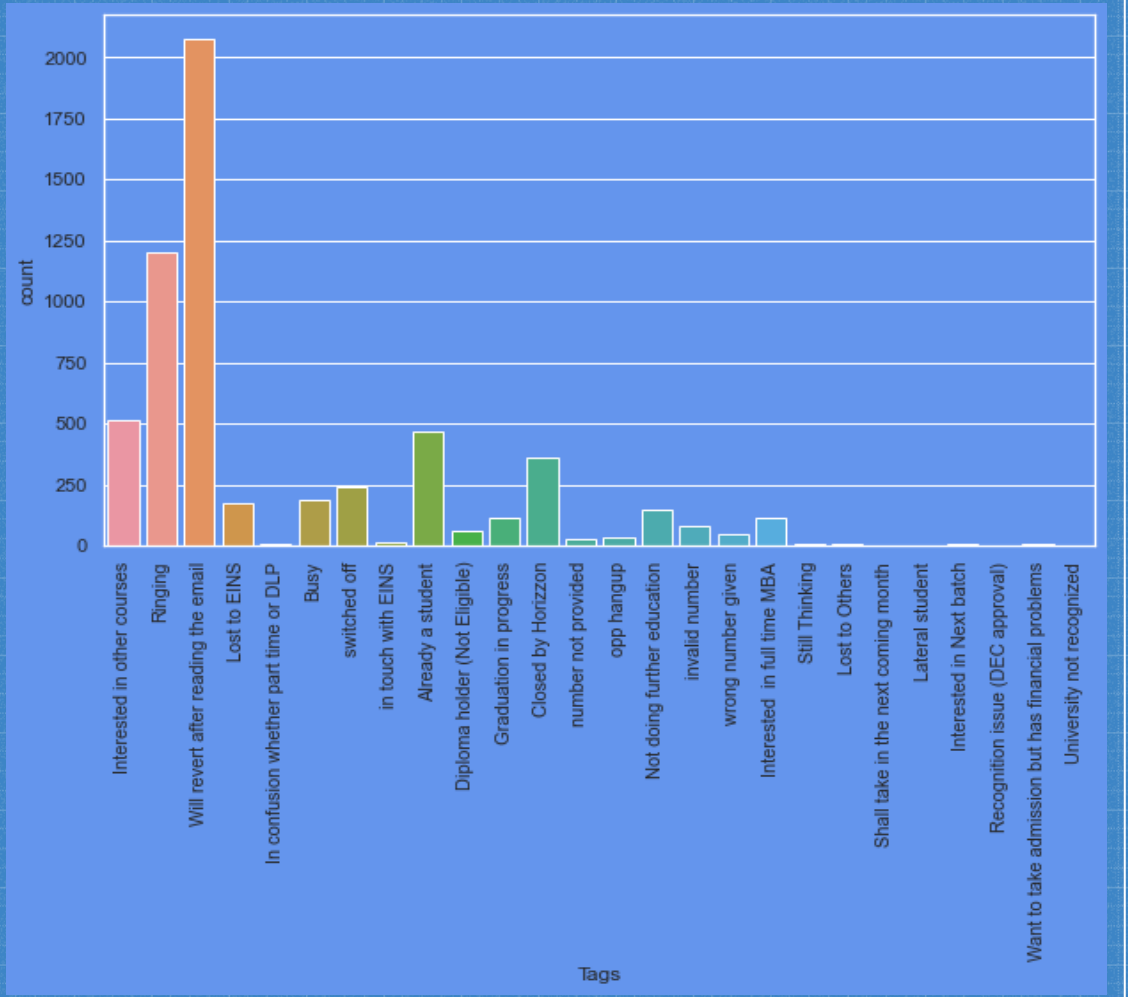
CITY

Since “Mumbai” is the highest occurrence in the data set, we replace NULL values with “Mumbai”



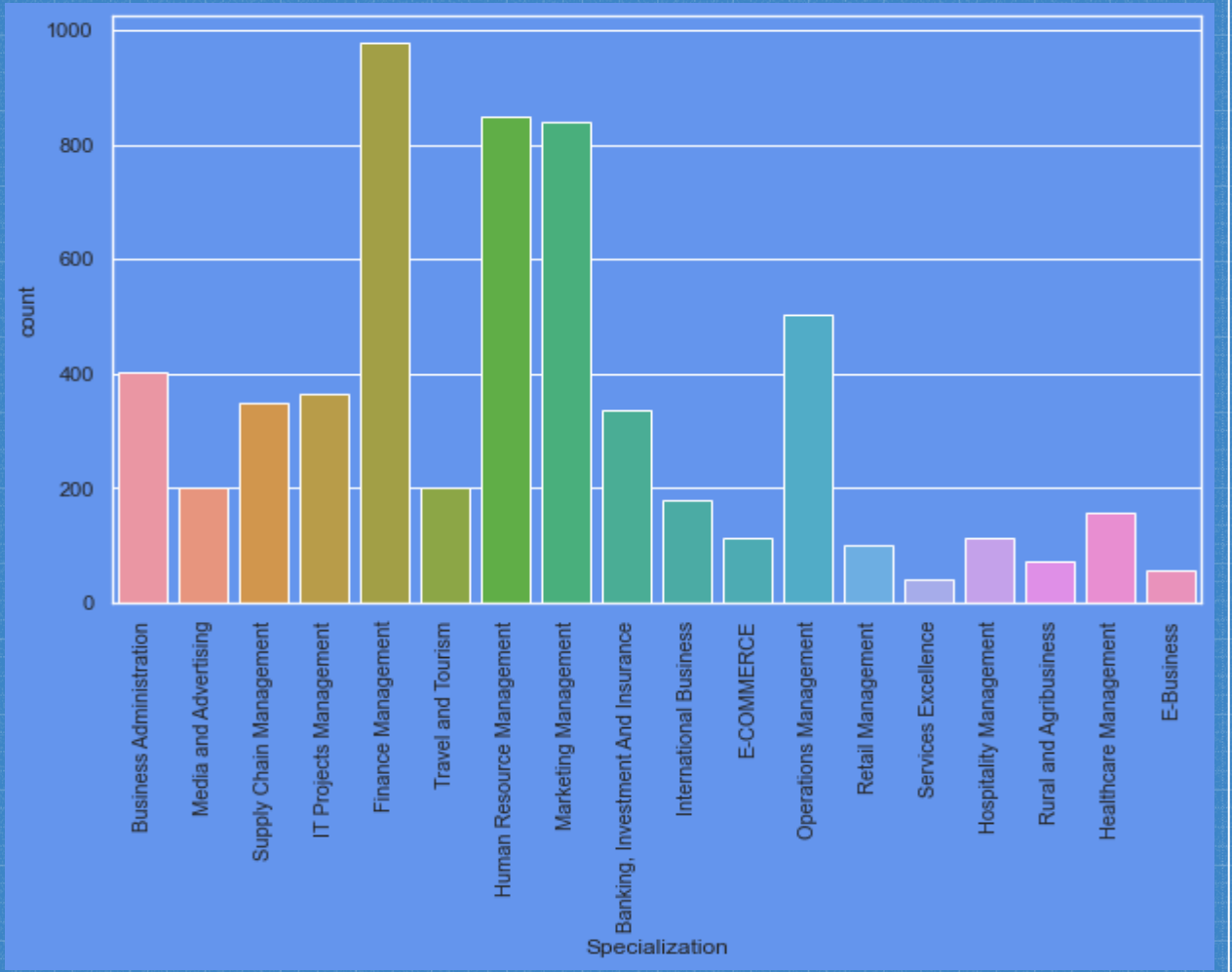
TAGS

“Will revert after reading the email” has the highest count in all the tags so we replace NULL values with “Will revert after reading the email”



SPECIALIZATION

"Finance Management" has the highest count but overall, not very high proportion so we replace NULL with "Others"



WHAT MATTERS THE MOST TO YOU IN CHOOSING A COURSE	WHAT IS YOUR CURRENT OCCUPATION	COUNTRY
“Better Career Prospects” has the highest count and, realistically speaking, that is why most people will join any course.	“Unemployed” has a very high count so we can safely replace NULL with “Unemployed”	“India” has very high count so we can safely replace NULL with “India”

TOTAL VISITS

Here 0 has the highest count but not by a large proportion so we replace NULL VALUES with mean

PAGE VIEWS PER VISIT

Here 0 has the highest count but not by a large proportion so we replace NULL VALUES with mean

LAST ACTIVITY

“Email Opened” has the highest count and since NULL values are only 1% we can replace them with “Email Opened”

LEAD SOURCE

“Google” is the highest count and NULL values are only 0.39 % so we can safely replace them with “Google”

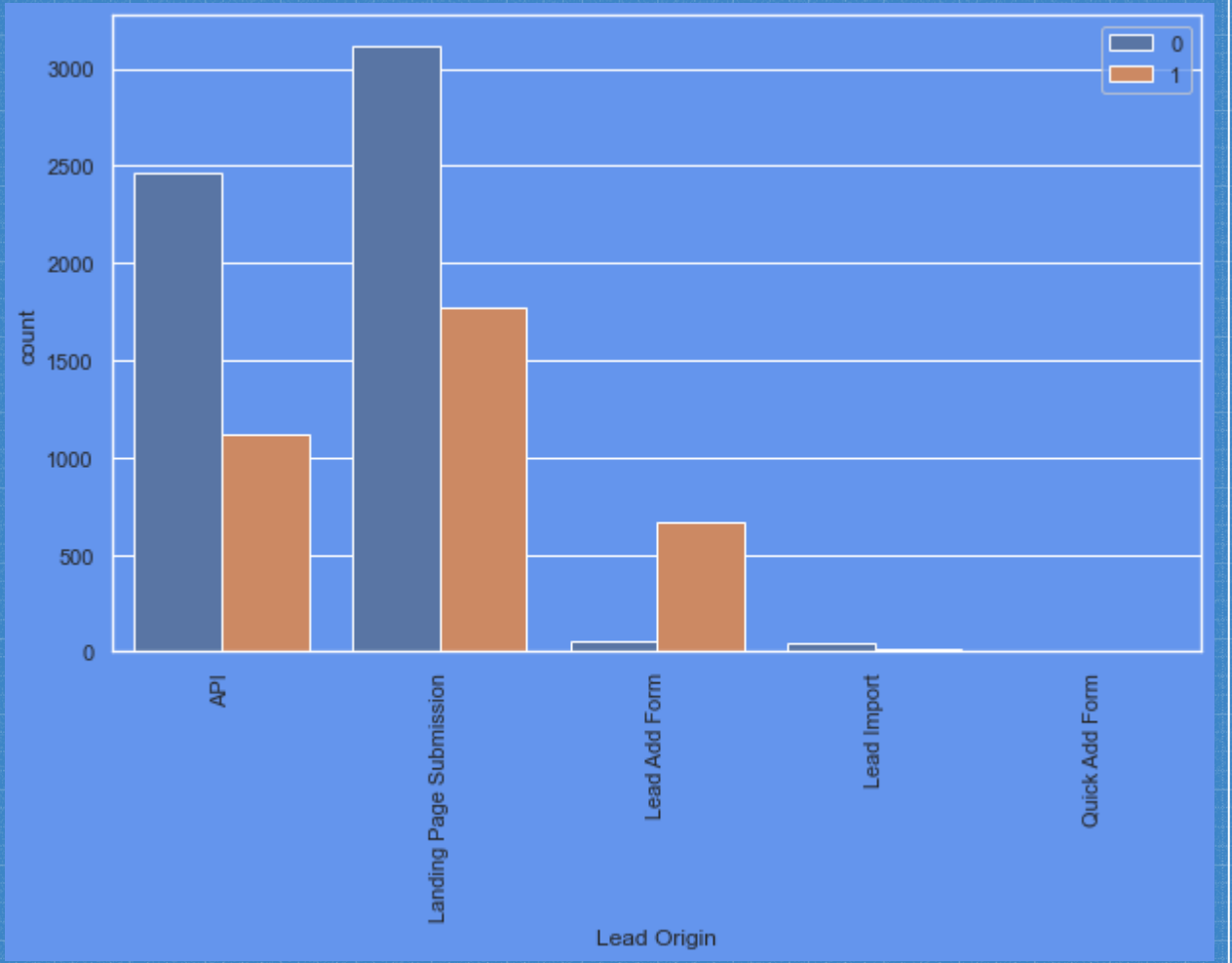
EXPLORATORY DATA ANALYSIS

UNIVARIATE ANALYSIS

LEAD ORIGIN

- API and Landing Page submission have approximately 40% and 56% conversion rate respectively and overall count from these two sources are high
- Lead Add Form has very high conversion rate, but overall conversion count is very low

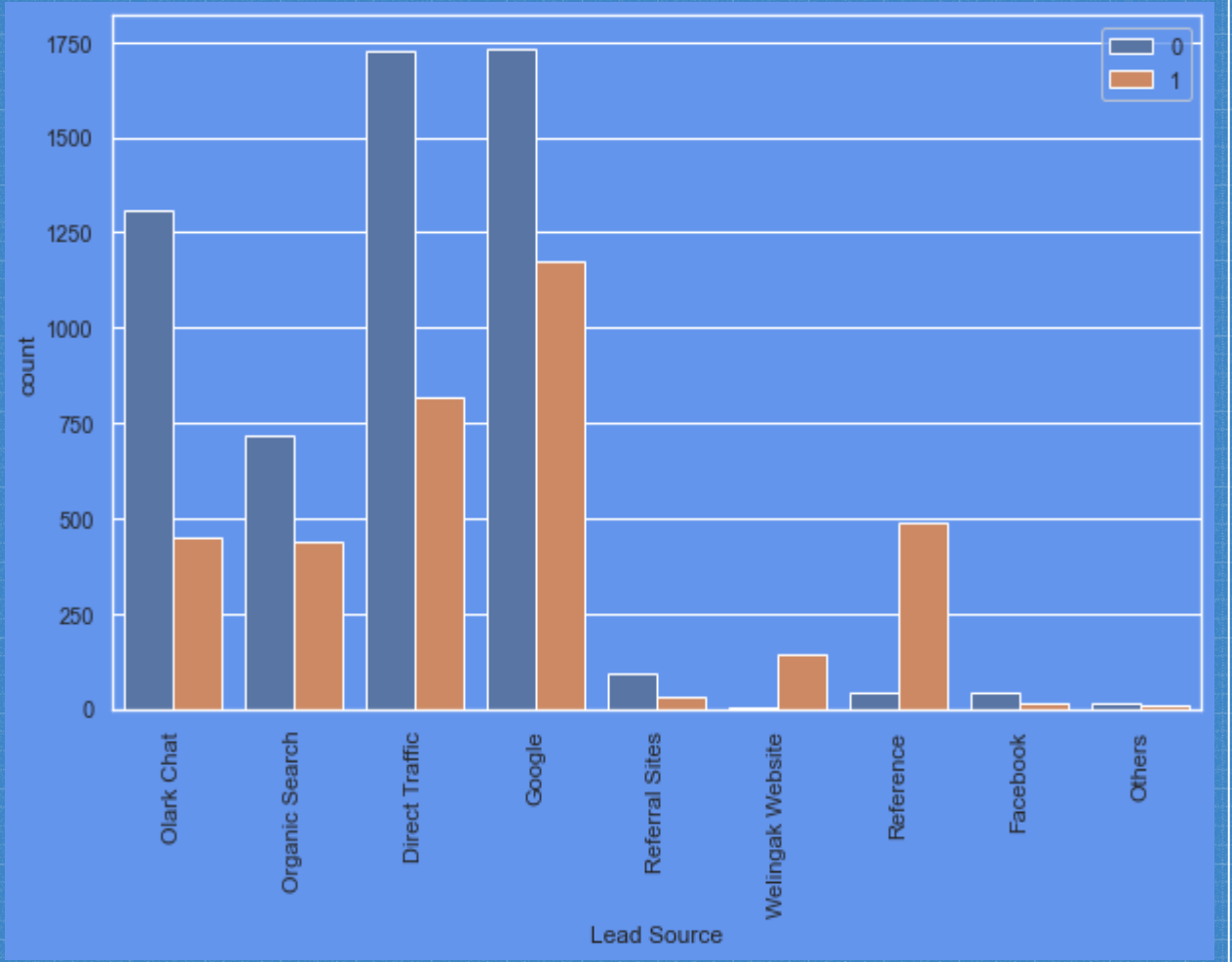
Since we have high conversion counts from API and Landing Page Submissions, we can focus on increasing the conversion rate from these two sources



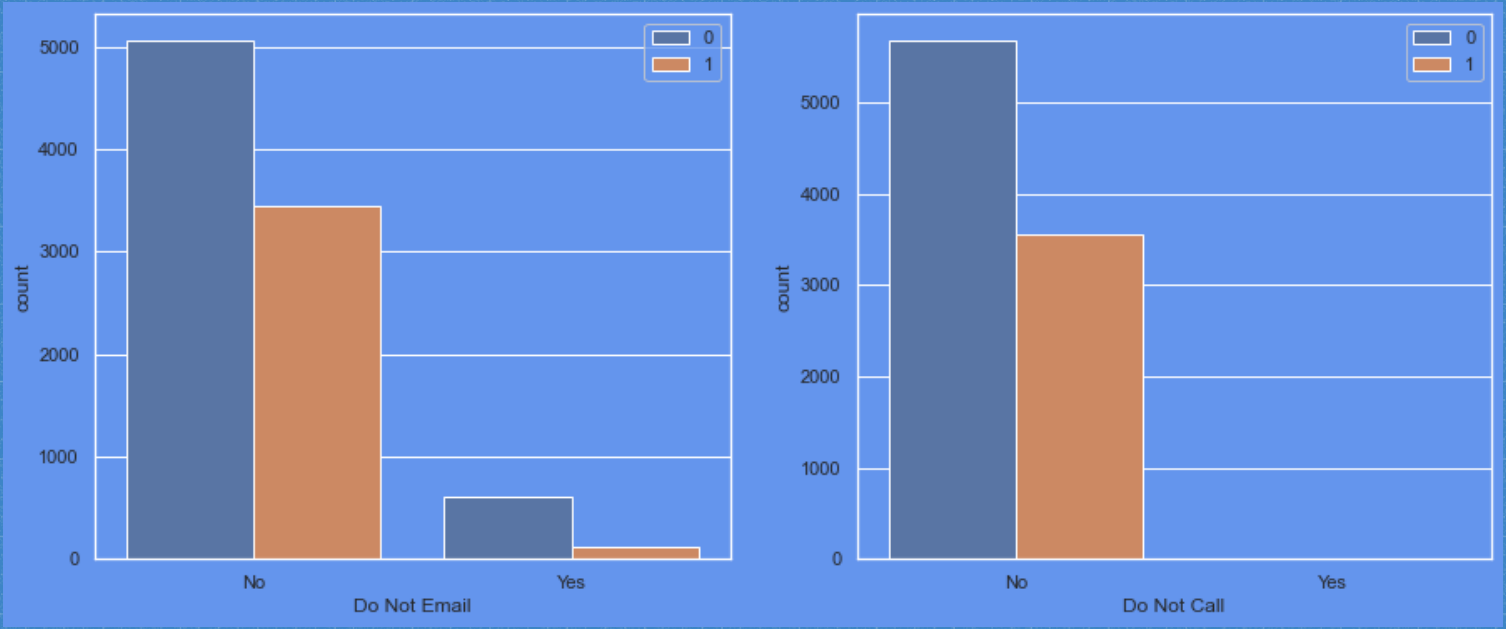
LEAD SOURCE

- Direct Traffic and Google have similar counts with Google having highest conversion rates
- Organic Search also has a relatively high conversion rate.
- Same goes for Reference but overall count is very less

To increase the overall conversion rate, we can focus on increasing the conversion rates from Google, Direct Traffic, Organic Search and Olark chat



DO NOT EMAIL & DO NOT CALL

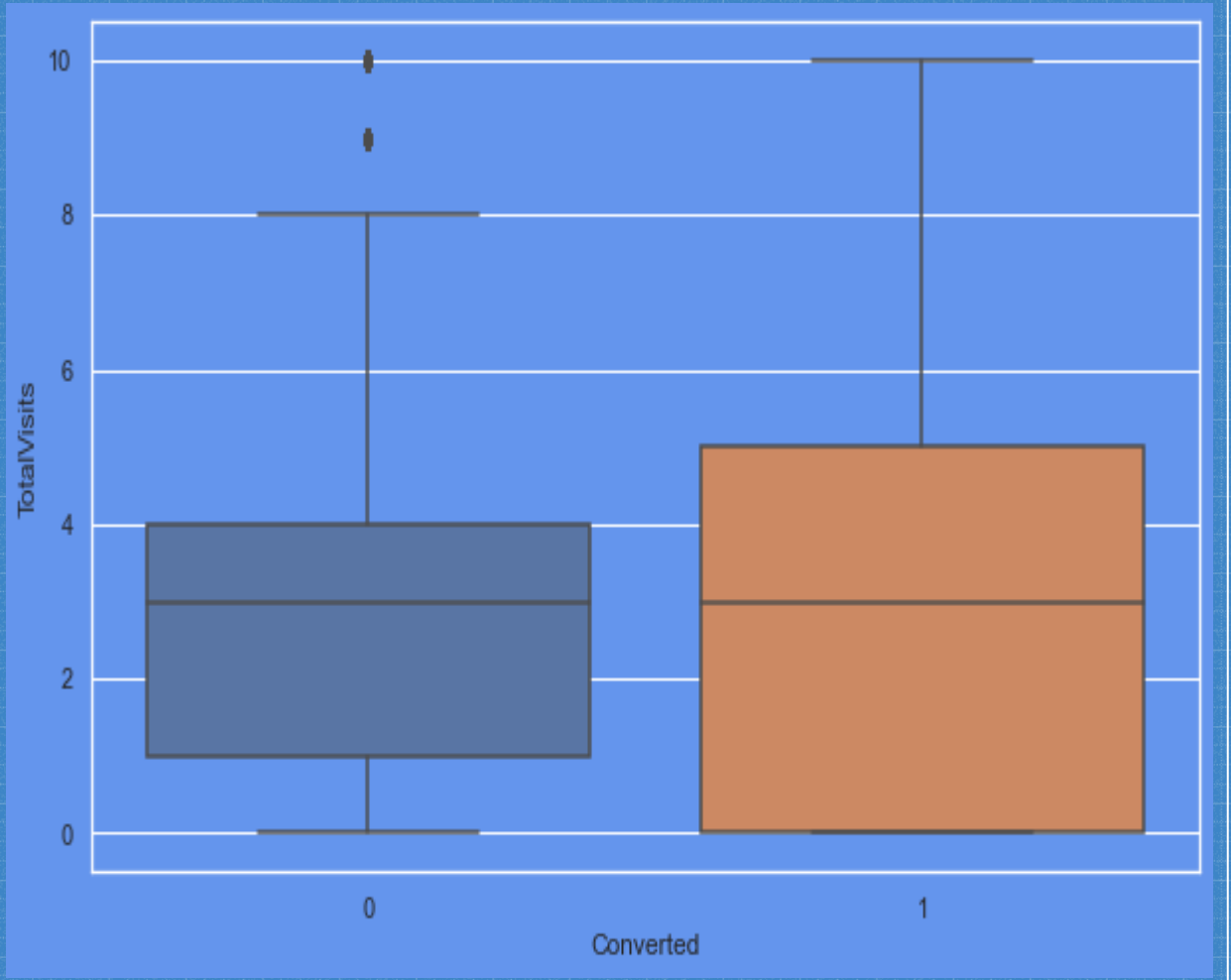


- People who said that they don't want to be Emailed have higher conversion rate than people who said that they wanted to be Emailed
- Same goes for the Do Not Call column as well

TOTAL VISITS

- Medians for not converted and converted are almost same
- people with 0 - 6 visits are seen to be converted but then again people with 1 - 4 visits are also seen to not be converted

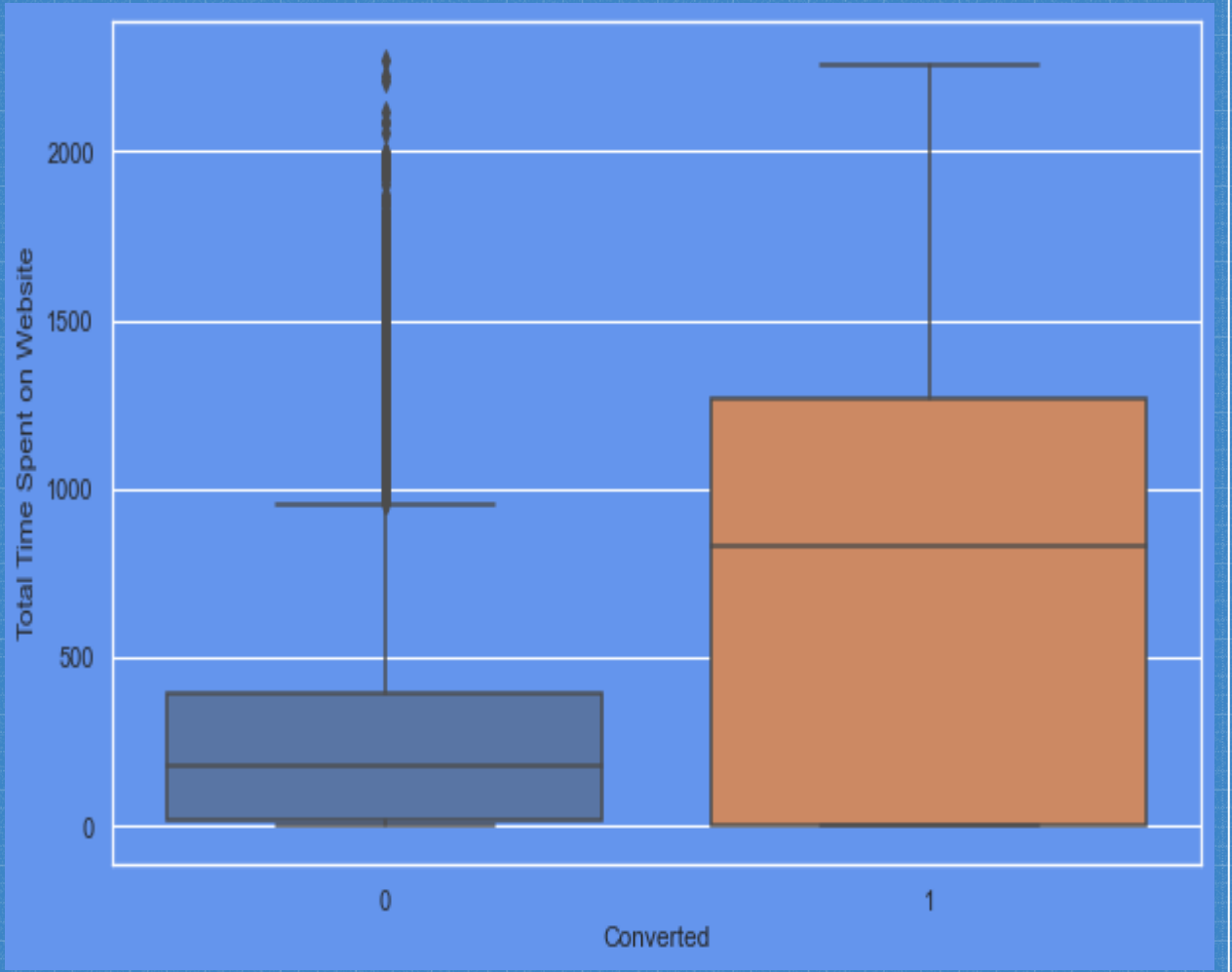
So, nothing conclusive is observed from this column



TOTAL TIME SPENT ON WEBSITE

- People spending more time on the website are more likely to be converted

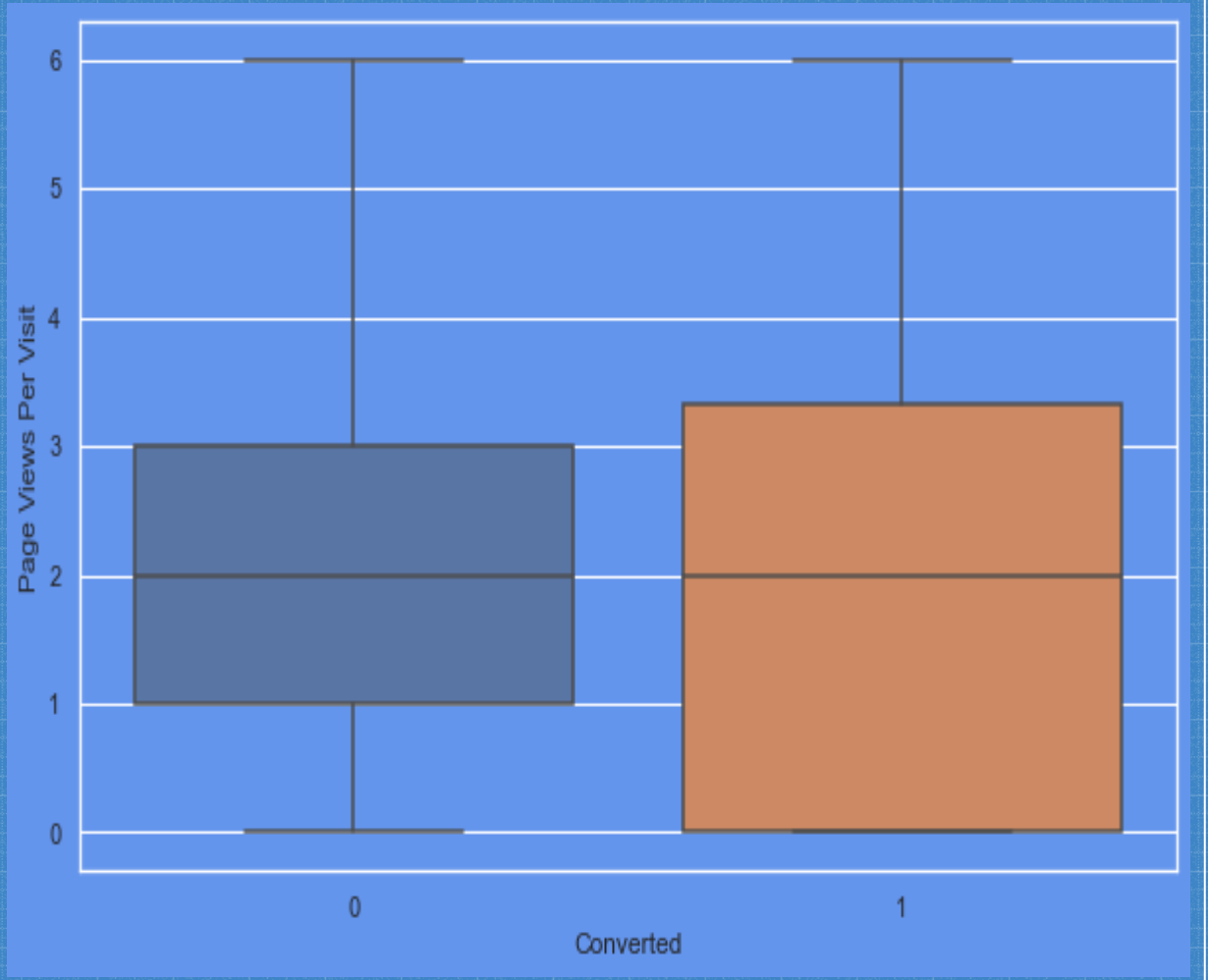
Keeping the website updated regularly is recommended



PAGE VIEWS PER VISIT

- Medians for both conversions and non conversions is same

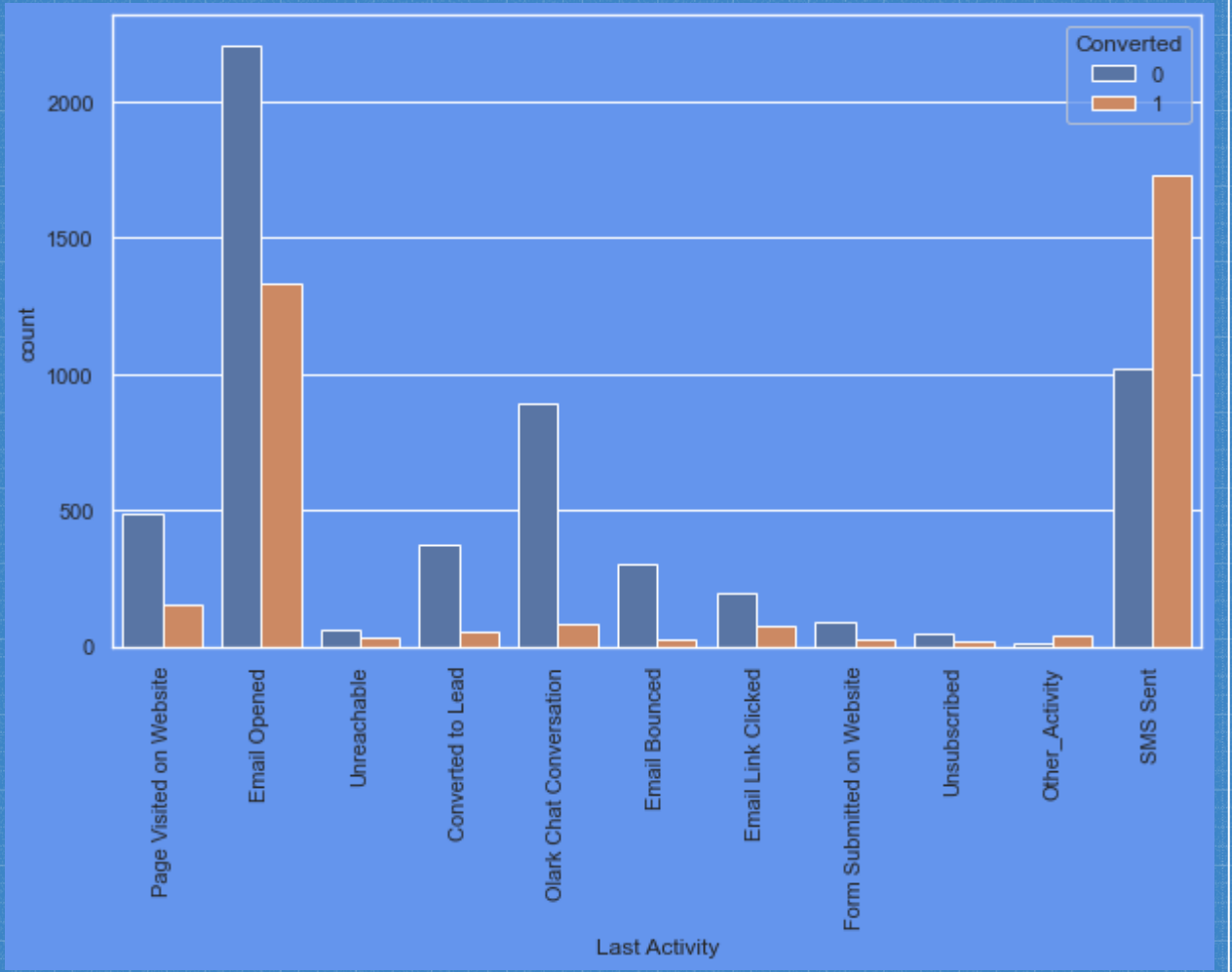
So, nothing conclusive can be said here



LAST ACTIVITY

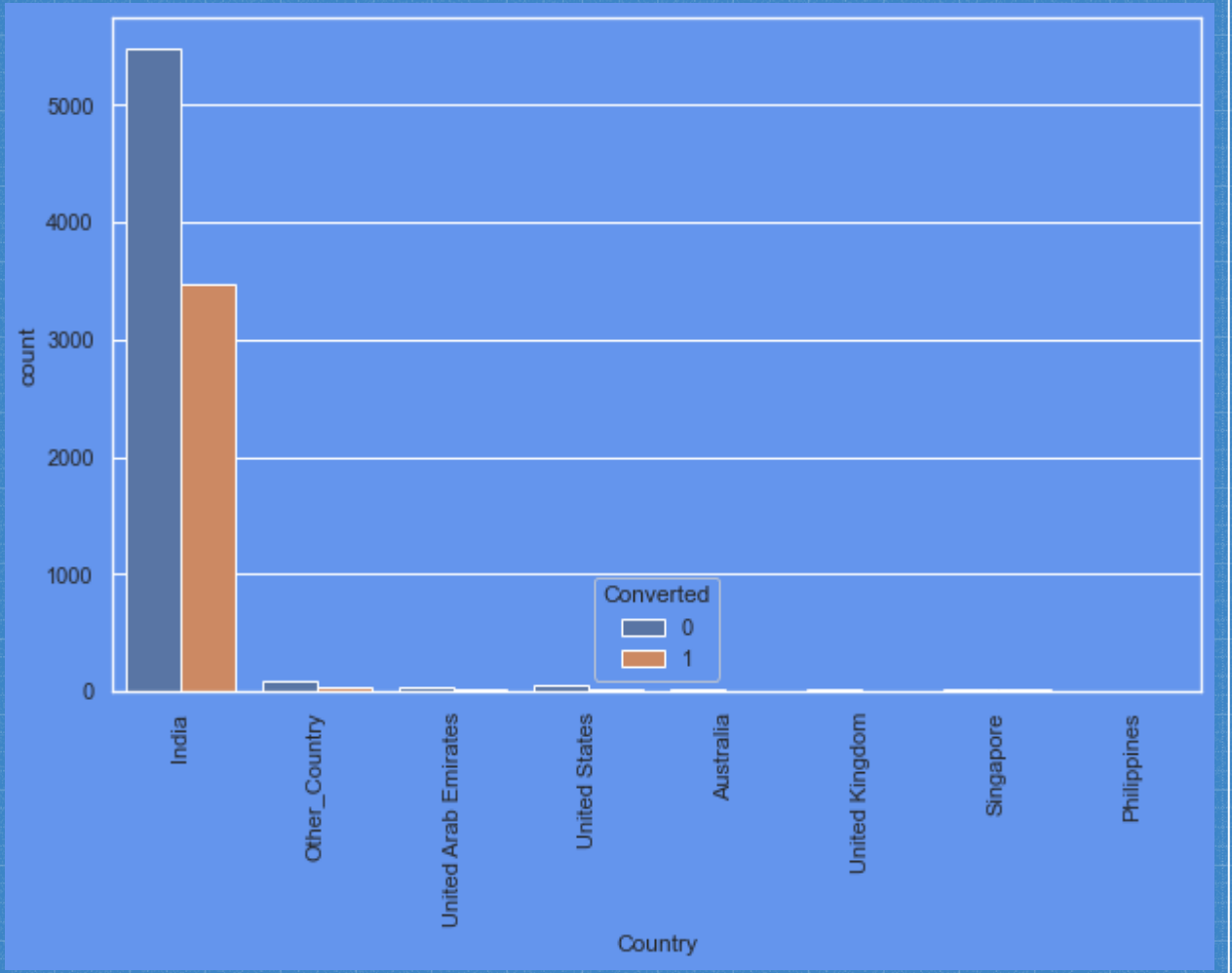
- People getting SMS have the highest conversion rate although their count is second highest, with people who are opening the emails having the highest count
- People having Olark char conversations are significant in number although their conversion rate is very low

Focus can be on increasing conversion rates for Email Opened, SMS sent, Olark chat conversations



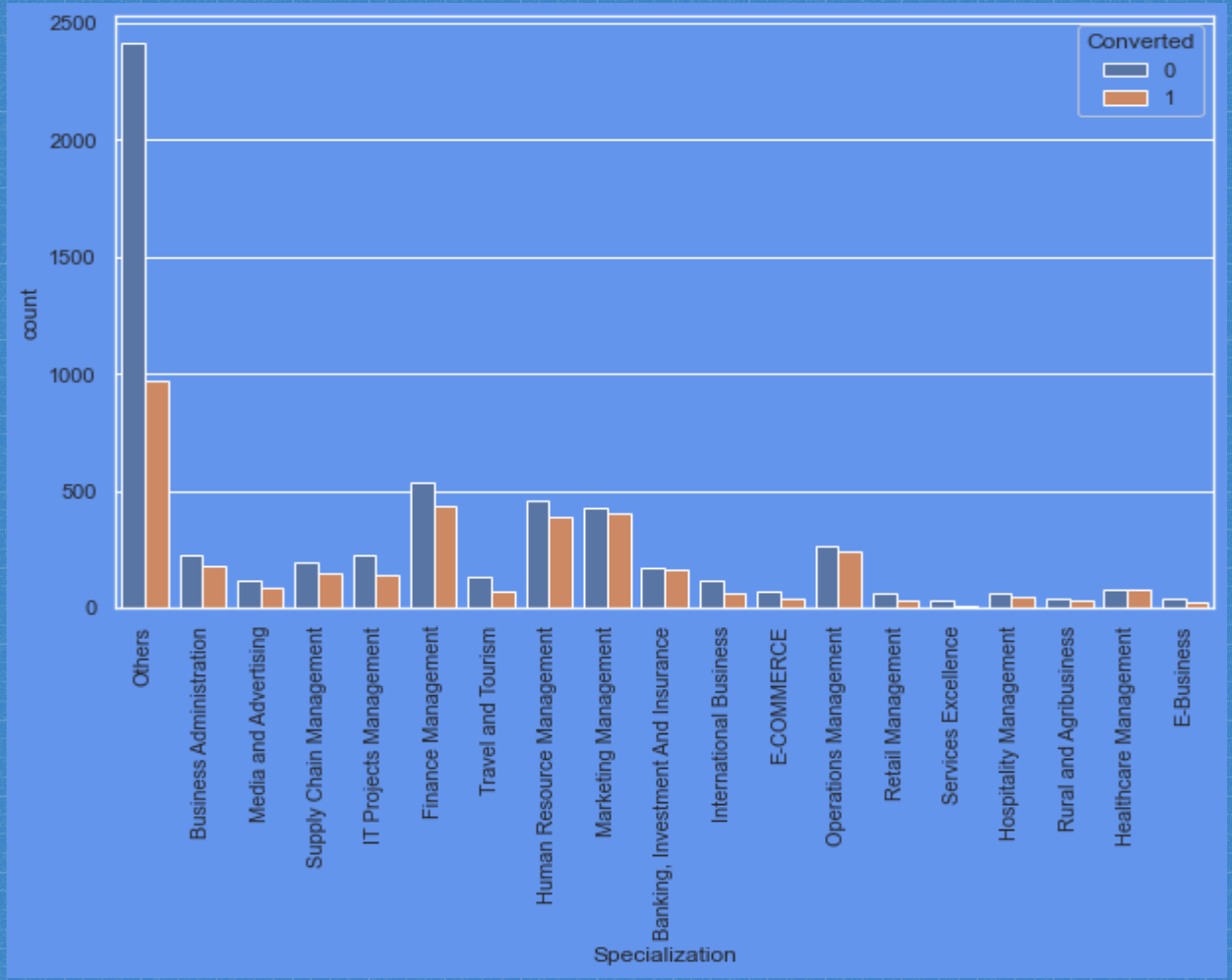
COUNTRY

Not much to conclude except the fact that India has the highest count



SPECIALIZATION

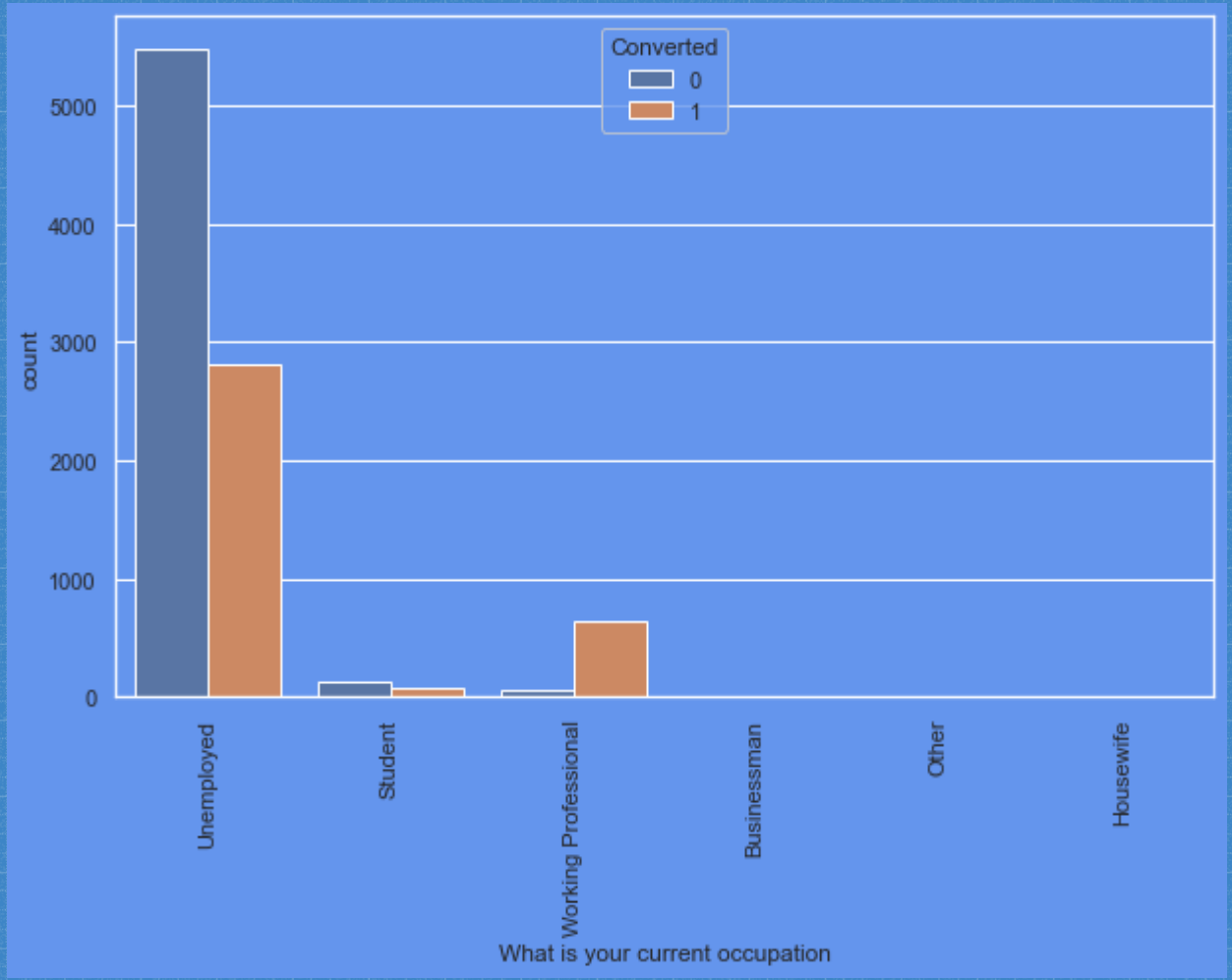
We need to focus on specializations having high conversion rates and try to increase them even further i.e., Finance Management, HR Management, Marketing Management, Operations Management etc.



WHAT IS YOUR OCCUPATION

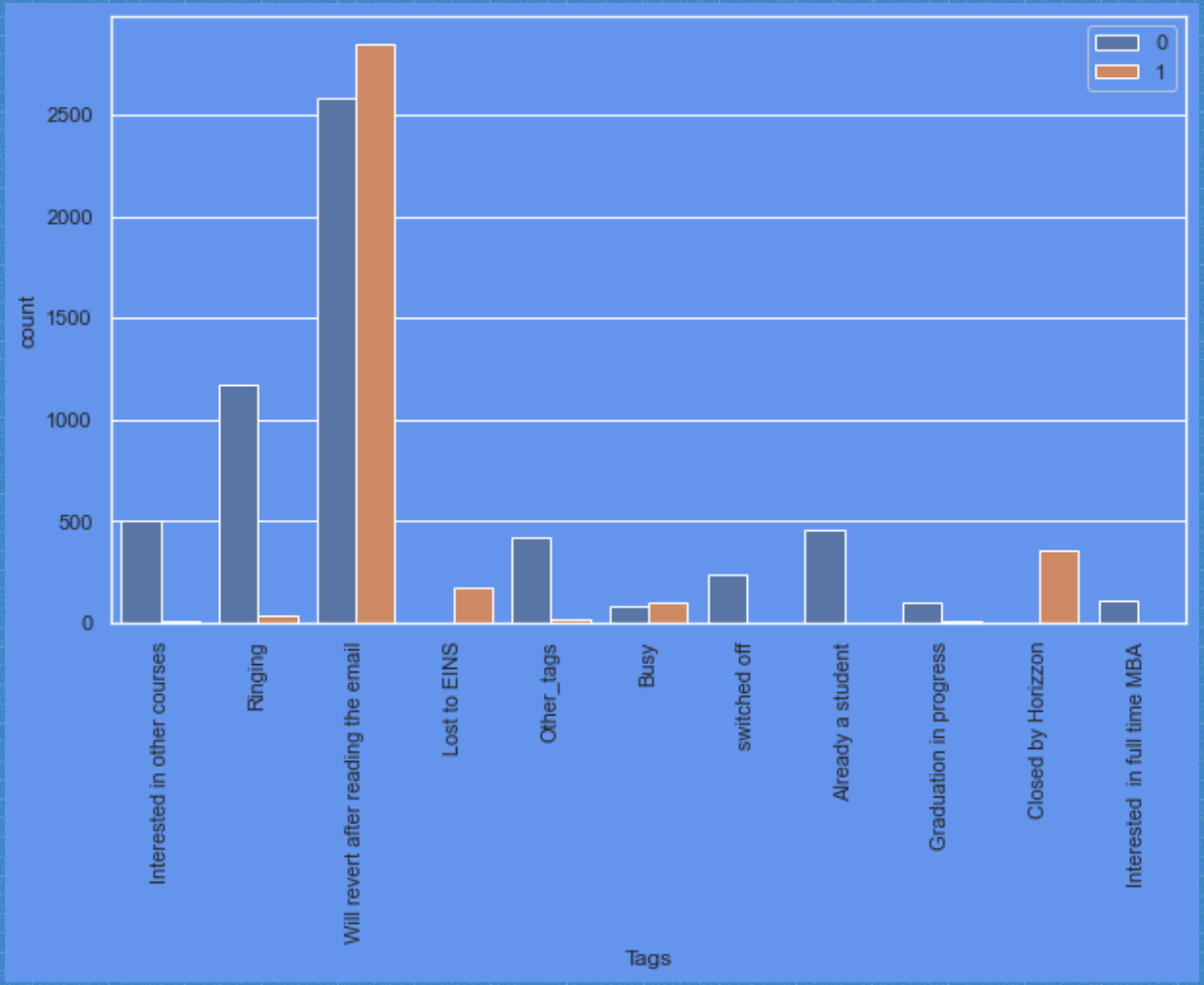
- Working professionals have high conversion rates although very less count
- Unemployed people, although high in number, have low conversion rate

Increasing the number of working professionals signing up and increasing the conversion rates of unemployed people will help



TAGS

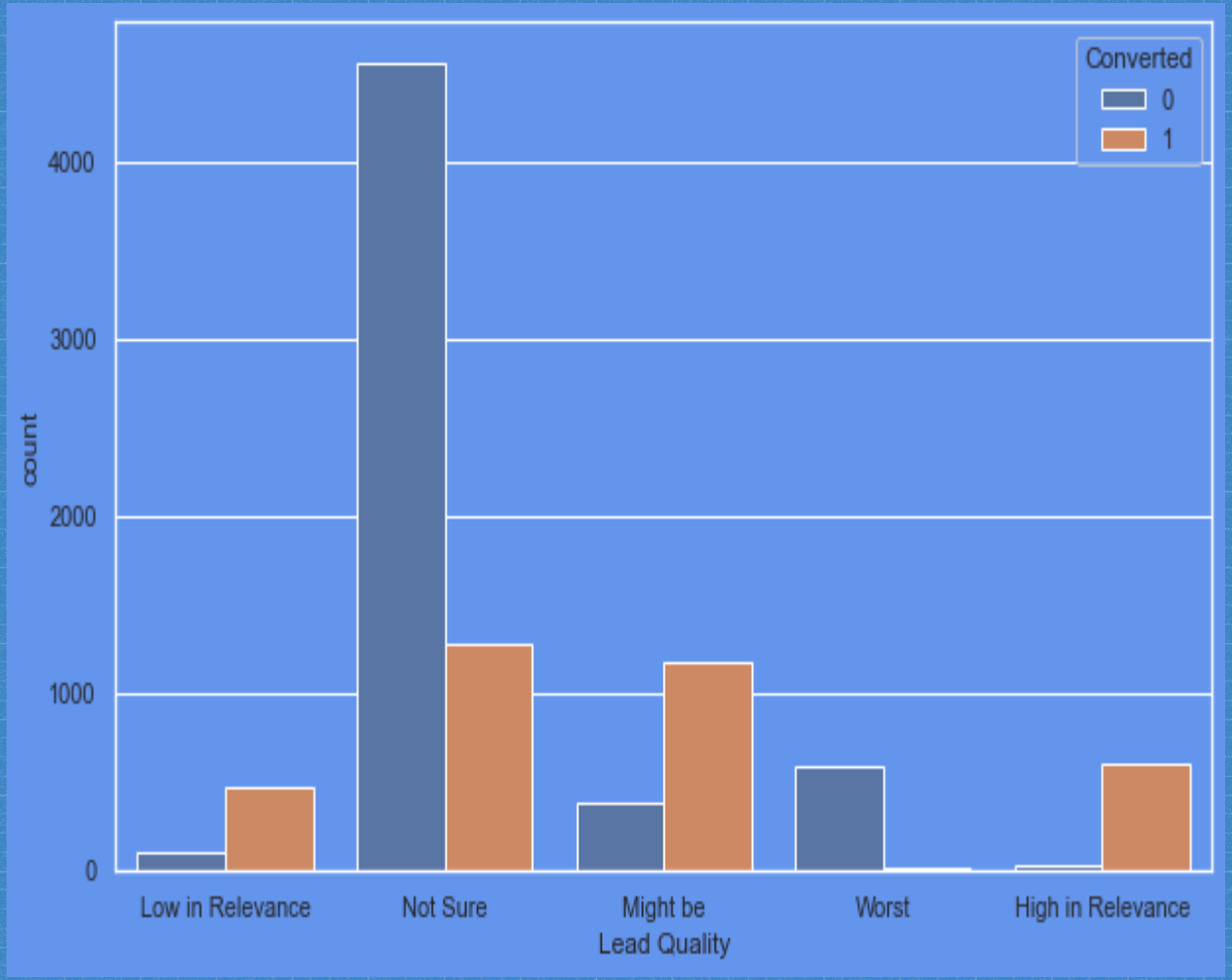
- People who say they will revert after reading the email have the highest conversion rates
- People who have been called and are not picking up are high in count but very low in terms of conversion rates. Same goes for people interested in other courses



LEAD QUALITY

- Max count is where the lead quality can't be determined thus the low conversion rate there.
- Proportionally speaking, the highest conversion rate is for the lead quality "High in Relevance" but its count is very less
- "Might be“ Lead Quality also has a high conversion rate

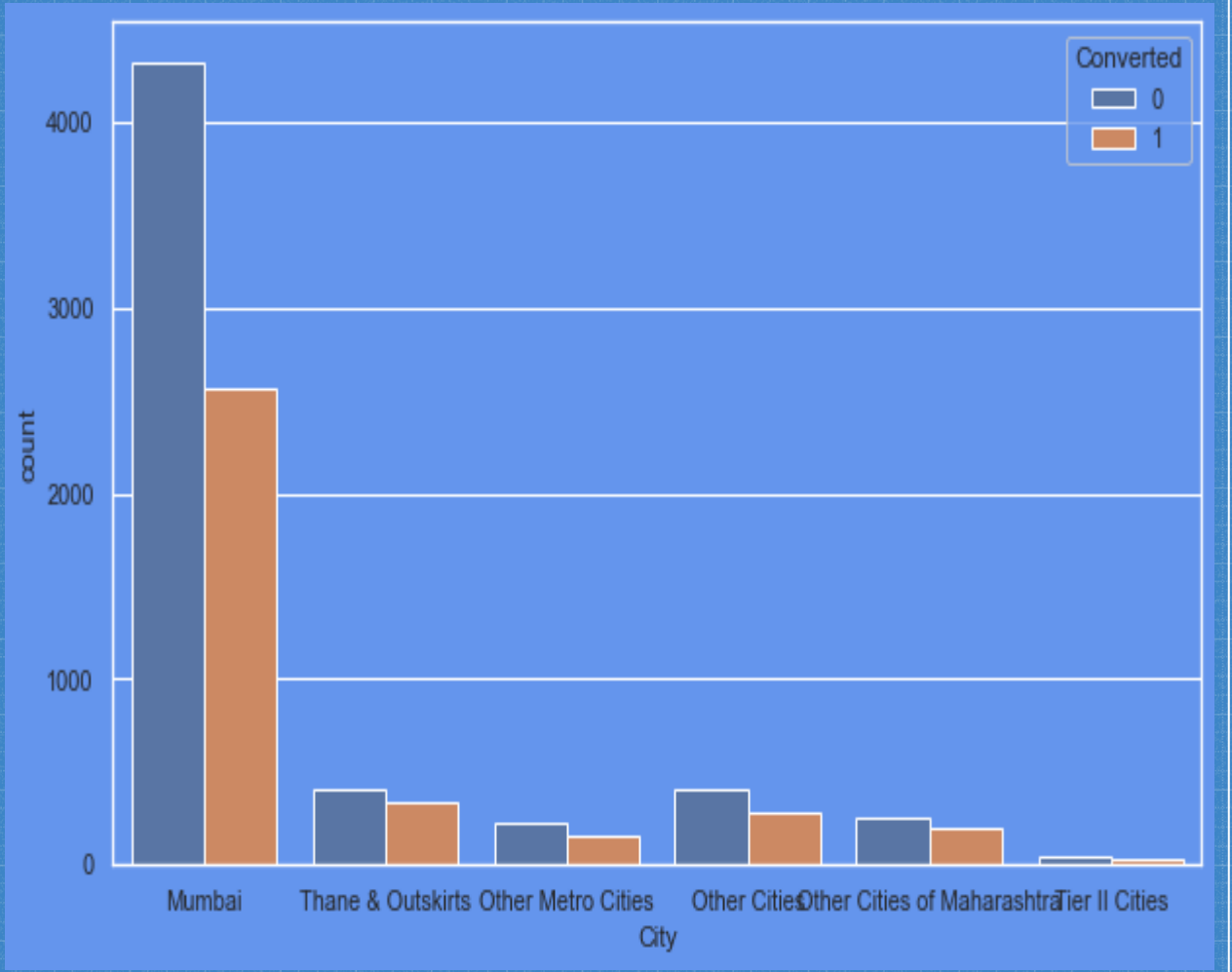
Getting high quality leads would be important



CITY

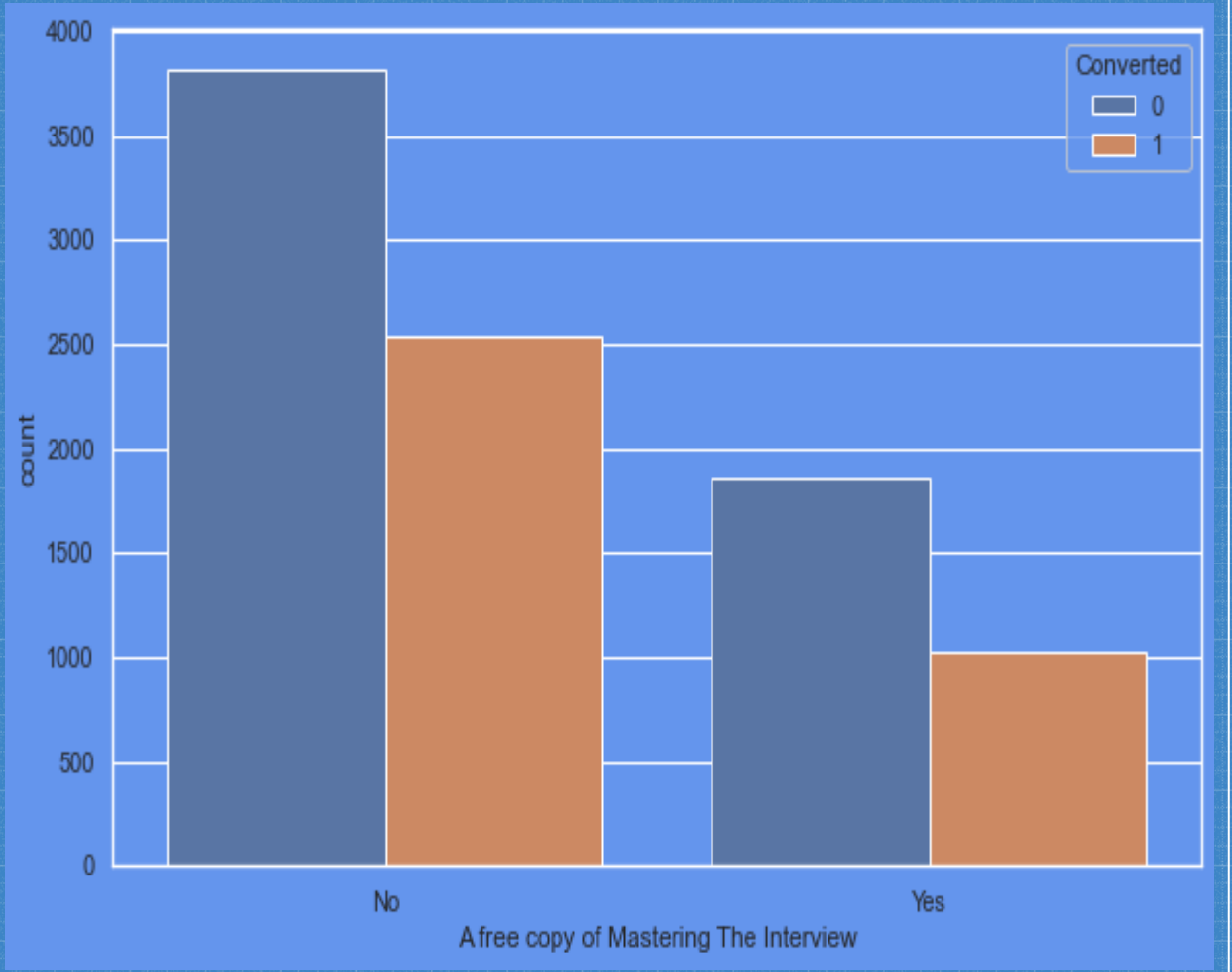
- Mumbai has the max count of people registering for courses and a decent conversion rate of around 50%
- Thane and outskirts has a higher conversion rate but very less count
- Same goes for other cities

Focus can be more people registering from Mumbai to increase their conversion rate



FREE COPY OF MASTERING THE INTERVIEW

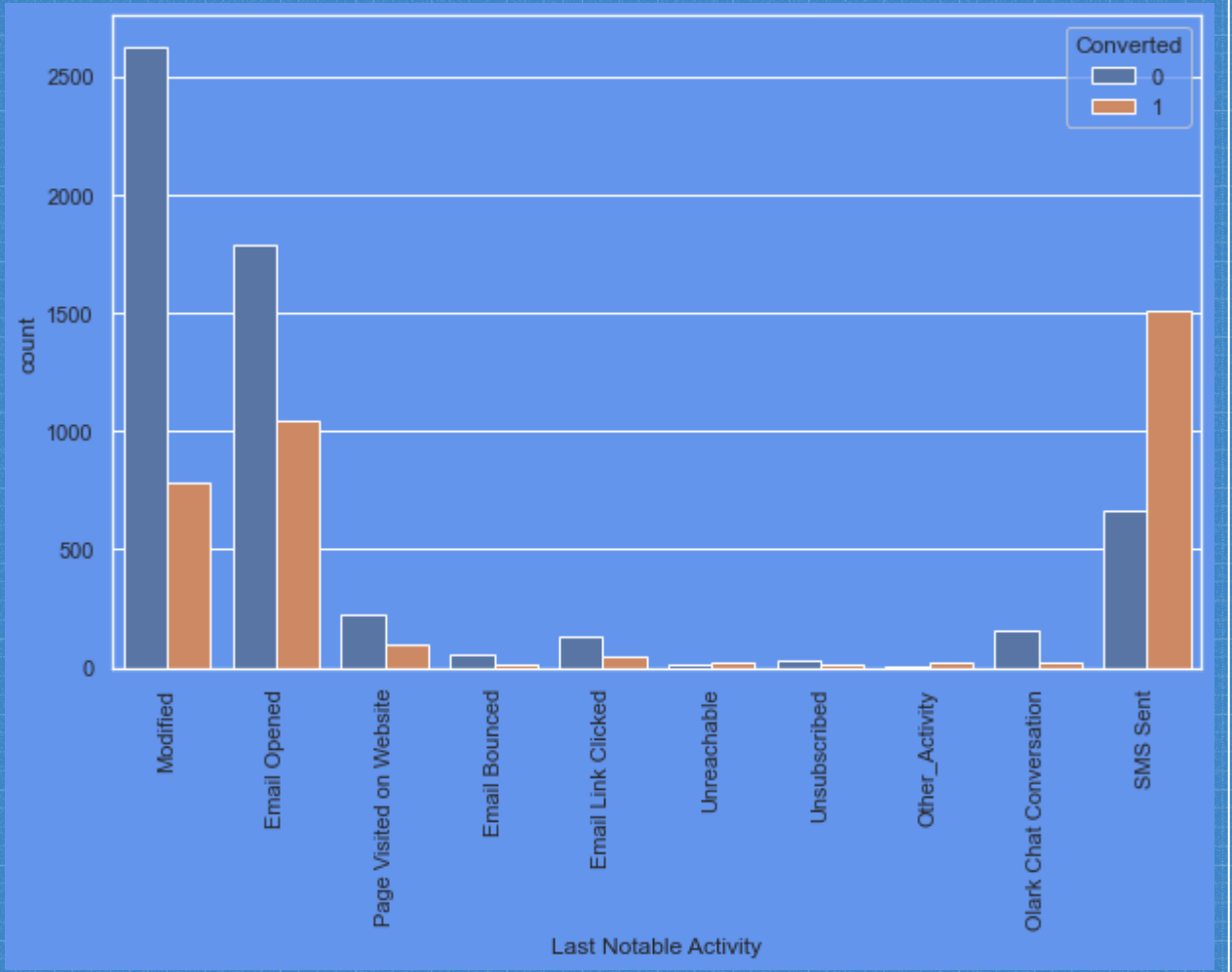
People who were not interested in getting a free copy of "Mastering the interview" have a higher conversion rate (and count) as compared to people who did opt for a free copy of "Mastering the interview"



LAST NOTABLE ACTIVITY

- "Modified" column might refer to people who might have modified their profile on website (just an assumption) and it has the highest count but very low conversion rate
- "SMS sent" has a high conversion rate but low count

Overall, this column will not really help us make a business decision



DATA PREPARATION

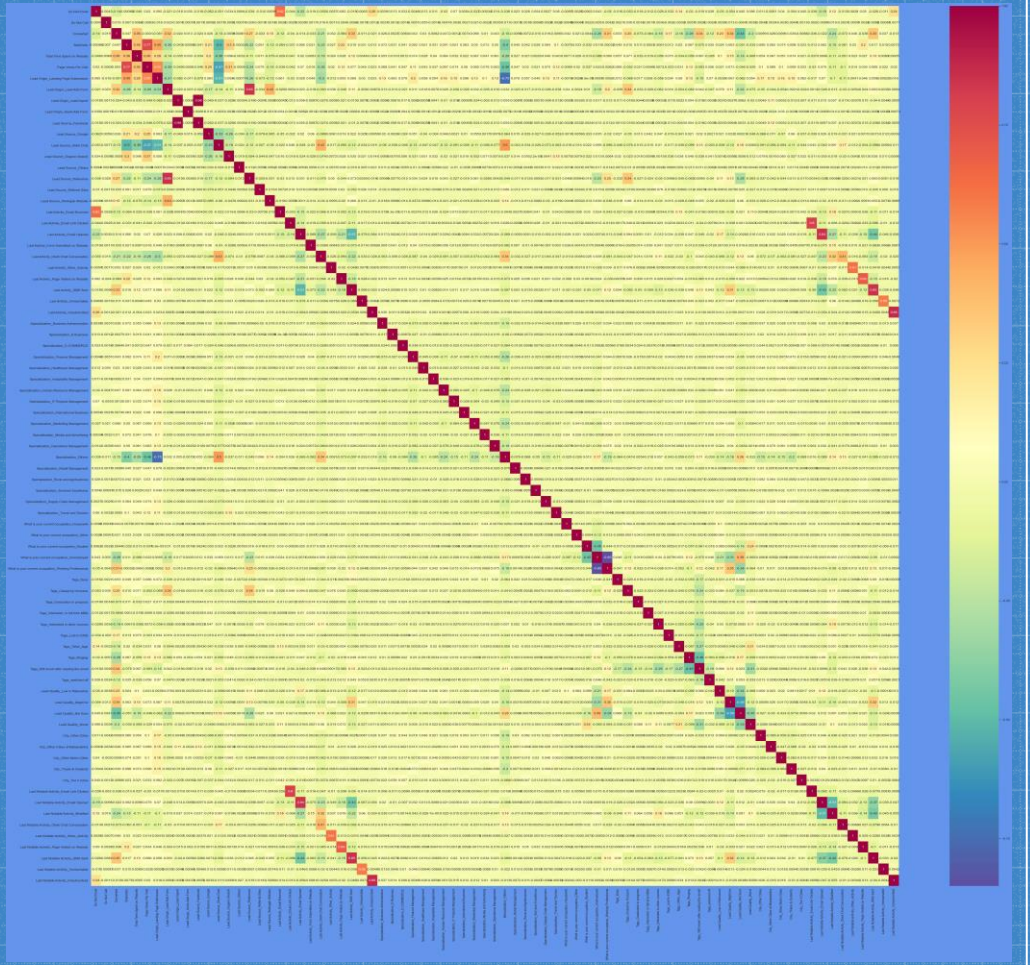
Activity	Columns on which it is to be performed
Converting Binary (Yes/No) variables to 1/0	• Do not email • Do not call
Creating dummies for categorical variables with multiple levels	• Lead Origin • Lead Source • Last Activity • Specialization • Current occupation • Tags • Lead Quality • City • Last Notable Activity

TRAIN TEST SPLIT AND FEATURE SCALING

- Define the X and y data sets
- Use the train_test_split method to split the data set into training and test data set (70-30 split)
  - Certain variables i.e., Total visits, total time spent on website and page views per visit are scaled using Standard Scaler

HEATMAP

- This was just a second measure to understand if we are missing any correlations which might help us with the model.
- There are not many high correlations except the ones like Last Activity_Unsubscribed and Last Notable Activity_Unsubscribed. These are the type of correlations that don't make sense as they are same variables told differently



COARSE TUNING USING RFE

- We use Recursive Feature Elimination to choose top 15 features

```
# Checking the top 15 columns
col_top_15 = X_train.columns[rfe.support_]
col_top_15
```

```
Index(['Lead Origin_Lead Add Form', 'Lead Source_Welingak Website',
       'Last Activity_SMS Sent',
       'What is your current occupation_Working Professional', 'Tags_Busy',
       'Tags_Closed by Horizzon', 'Tags_Lost to EINS', 'Tags_Ringing',
       'Tags_Will revert after reading the email', 'Tags_switched off',
       'Lead Quality_Not Sure', 'Lead Quality_Worst',
       'Last Notable Activity_Modified',
       'Last Notable Activity_Olark Chat Conversation',
       'Last Notable Activity_SMS Sent'],
      dtype='object')
```

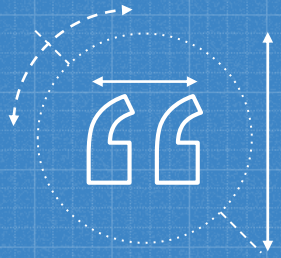
FINAL MODEL

```
# Logistic regression model with top 15 columns chosen
X_train_sm = sm.add_constant(X_train[col_top_15])

logis_model_2 = sm.GLM(y_train, X_train_sm, family=sm.families.Binomial())

res = logis_model_2.fit()

res.summary()
```

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6468
Model:	GLM	**Df Residuals:**	6452
Model Family:	Binomial	**Df Model:**	15
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1586.9
Date:	Thu, 08 Apr 2021	**Deviance:**	3173.8
Time:	22:32:57	**Pearson chi2:**	3.39e+04
No. Iterations:	9		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.4965	0.214	-7.007	0.000	-1.915	-1.078
Lead Origin_Lead Add Form	0.8900	0.298	2.986	0.003	0.306	1.474
Lead Source_Welingak Website	3.3981	0.795	4.272	0.000	1.839	4.957
Last Activity_SMS Sent	1.2192	0.183	6.678	0.000	0.861	1.577
What is your current occupation_Working Professional	1.3295	0.281	4.731	0.000	0.779	1.880
Tags_Busy	3.5753	0.315	11.337	0.000	2.957	4.193
Tags_Closed by Horizzon	9.1344	1.045	8.739	0.000	7.086	11.183
Tags_Lost to EINS	9.3595	0.754	12.421	0.000	7.883	10.836
Tags_Ringing	-1.7636	0.313	-5.635	0.000	-2.377	-1.150
Tags_Will revert after reading the email	3.7308	0.222	16.769	0.000	3.295	4.167
Tags_switched off	-2.5007	0.578	-4.327	0.000	-3.633	-1.368
Lead Quality_Not Sure	-3.3366	0.131	-25.433	0.000	-3.594	-3.079
Lead Quality_Worst	-3.4673	0.684	-5.070	0.000	-4.808	-2.127
Last Notable Activity_Modified	-1.3000	0.124	-10.499	0.000	-1.543	-1.057
Last Notable Activity_Olark Chat Conversation	-1.2259	0.362	-3.384	0.001	-1.936	-0.516
Last Notable Activity_SMS Sent	1.0226	0.219	4.675	0.000	0.594	1.451



Training model accuracy is 91.91%

SENSITIVITY, SPECIFICITY AND OTHER METRICS

Metric	Value
Sensitivity	0.85
Specificity	0.96
False Positive Rate	0.04
Positive Predictive Value	0.93
Negative Predictive Value	0.91
True Positive Rate	0.85
False Positive Rate	0.04

VIF

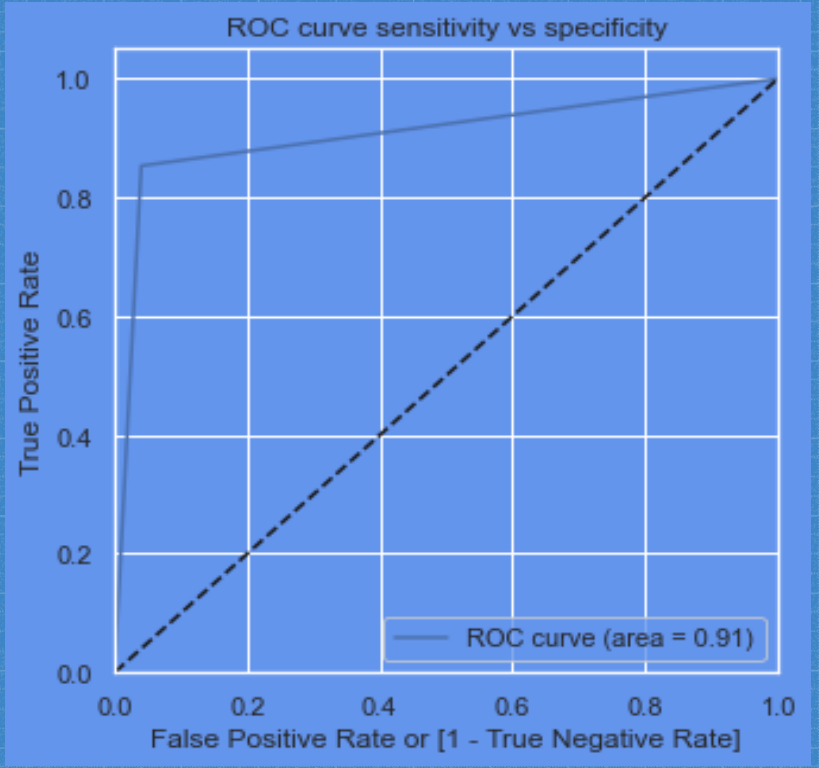
- All features have VIF less than 2, so we don’t need to eliminate any feature

	Features	VIF
5	Tags_Closed by Horizzon	1.28
1	Lead Source_Welingak Website	1.27
9	Tags_switched off	1.12
4	Tags_Busy	1.11
6	Tags_Lost to EINS	1.06
13	Last Notable Activity_Olark Chat Conversation	1.05
3	What is your current occupation_Working Profes...	0.63
2	Last Activity_SMS Sent	0.46
11	Lead Quality_Worst	0.43
8	Tags_Will revert after reading the email	0.15
7	Tags_Ringing	0.14
12	Last Notable Activity_Modified	0.07
10	Lead Quality_Not Sure	0.02
0	Lead Origin_Lead Add Form	0.01
14	Last Notable Activity_SMS Sent	0.00

PLOTTING ROC CURVE

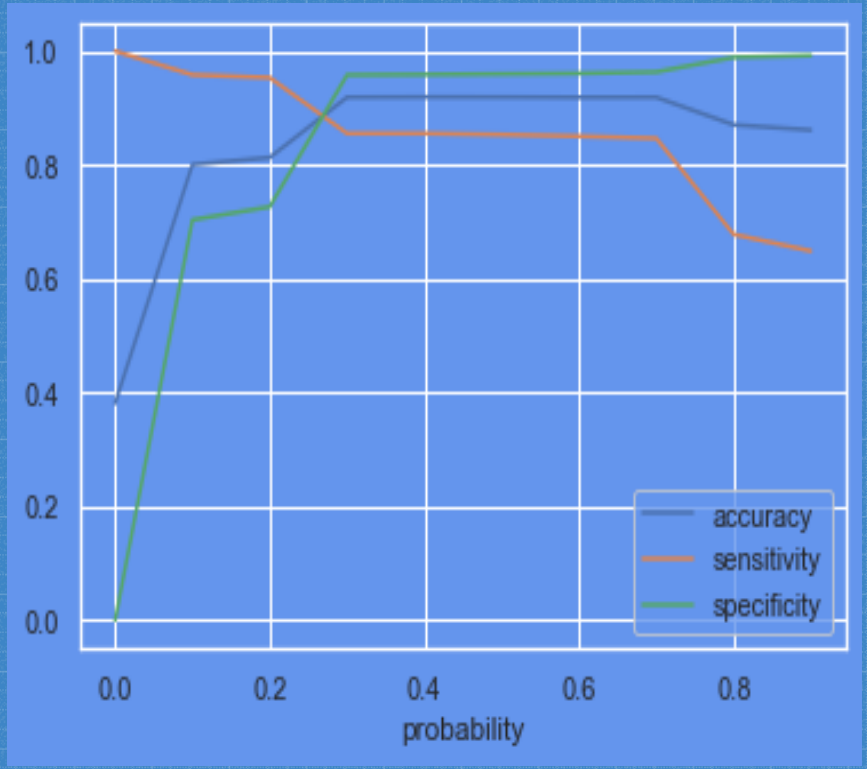
An ROC curve will help us understand the below things:

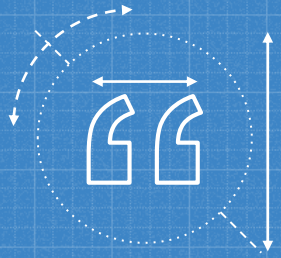
- It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will cause a decrease in specificity).
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.



FINDING OPTIMAL CUT-OFF POINT

- From the above curve **0.233** is the optimum probability as that's where the accuracy, sensitivity and specificity coincide





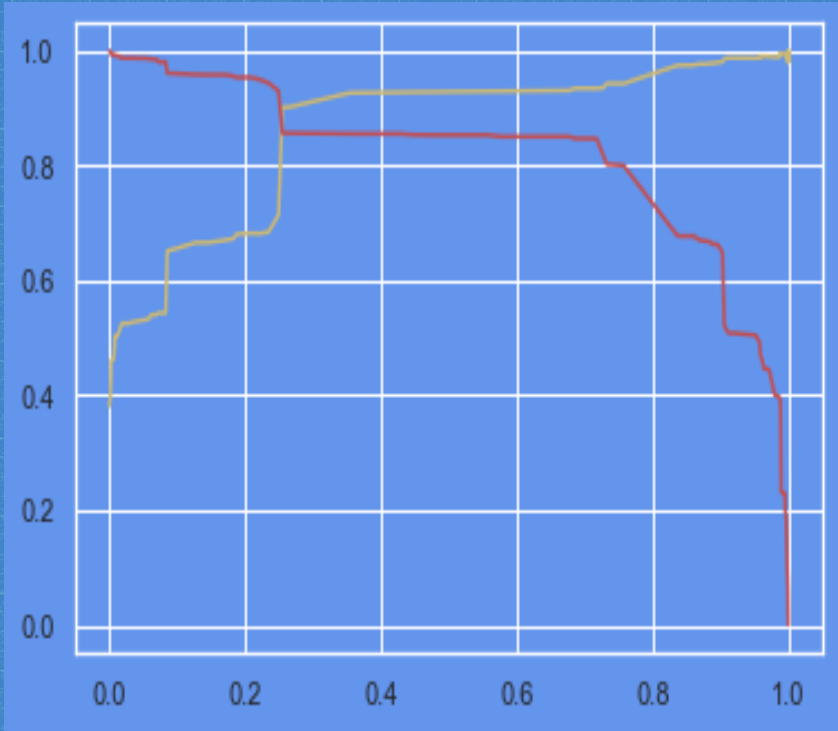
Test model accuracy is 81.5%

SENSITIVITY, SPECIFICITY AND OTHER METRICS FOR TEST SET

Metric	Value
Sensitivity	0.95
Specificity	0.72
False Positive Rate	0.28
Positive Predictive Value	0.69
Negative Predictive Value	0.96
True Positive Rate	0.95
False Positive Rate	0.28

PRECISION AND RECALL TRADE-OFF

- Precision – 0.68
- Recall – 0.94



FINAL RECOMMENDATIONS

FINAL RECOMMENDATIONS

- People count needs to be increased for Lead_Add Form, Lead Source_Wellingak website, Last Activity_SMS sent, Tags_Lost to EINS, Tags_Closed by Horizzon
- Keep the website regularly updated
- Better record keeping about specializations of the people
- More Working Professionals need to be brought on the platform
- People who say they will revert after reading the email, need to be nurtured by constant contact

- No need to contact people who are not responding and switched off
- No need to contact people for whom, the lead quality is Not Sure and Worst
- No need to contact people whose last activity is Olark chat conversation and last notable activity is SMS sent